On the Information Loss in Memoryless Systems: The Multivariate Case

Bernhard C. Geiger*, Gernot Kubin*
*Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
{geiger,gernot.kubin}@tugraz.at

***Abstract*—In this work we give a concise definition of information loss from a system-theoretic point of view. Based on this definition, we analyze the information loss in memoryless input-output systems subject to a continuous-valued input. For a certain class of multiple-input, multiple-output systems the information loss is quantified. An interpretation of this loss is accompanied by upper bounds which are simple to evaluate.**

Finally, a class of systems is identified for which the information loss is necessarily infinite. Quantizers and limiters are shown to belong to this class.

I. Introduction

In the XXXI. Shannon lecture Han argued that information theory links information-theoretic quantities, such as entropy and mutual information, to operational quantities such as source, channel, capacity, and error probability [1]. In this work we try to make a new link to an operational quantity not mentioned by Han: information loss. Information can be lost, on the one hand, in erasures or due to superposition of noise as it is known from communication theory. Dating back to Shannon [2] this loss is linked to the conditional entropy of the input given the output, at least in discrete-amplitude, memoryless settings. On the other hand, as stated by the data processing inequality (DPI, [3]), information can be lost in deterministic, noiseless systems. It is this kind of loss that we will treat in this work, and we will show that it makes sense to link it to the same information-theoretic quantity.

The information loss in input-output systems is very sparsely covered in the literature. Aside from the DPI for discrete random variables (RV) and static systems, some results are available for jointly stationary stochastic processes [4]. Yet, all these results just state that *information is lost*, without quantifying this loss. Only in [5] the information lost by collapsing states of a discrete-valued stochastic process is quantified as the difference between the entropy rates at the input and the output of the memoryless system.

Conversely, energy loss in input-output systems has been deeply analyzed, leading to meaningful definitions of transfer functions and notions of passivity, stability, and losslessness. Essentially, it is our aim to develop a system theory not from an energetic, but from an information-theoretic point of view. So far we analyzed the information loss of discrete-valued stationary stochastic processes in finite-dimensional dynamical input-output systems [6], where we proposed an upper bound on the information loss and identified a class of information-preserving systems (the information-theoretic counterpart to lossless systems). In [7] the information loss of continuous RVs in memoryless systems was quantified and bounded in a preliminary way. In this work, extending [7], we analyze the information loss for static multiple-input, multiple-output systems which are subject to a continuous input RV. Unlike in our previous work, we permit functions which lose an infinite amount of information and present the according conditions. Aside from that we provide a link between information loss and differential entropy, a quantity which is not invariant under changes of variables. The next steps towards an information-centered system theory are the analysis of discrete-time dynamical systems with continuous-valued stationary input processes and a treatment of information loss in multirate systems.

In the remainder of this paper we give a mathematically concise definition of information loss (Section II). After restricting the class of systems in Section III, in Section IV we provide exact results for information loss together with simple bounds, and establish a link to differential entropies. Finally, in Section V we show under which conditions the information loss becomes infinite.

This manuscript is an extended version of a paper submitted to a conference.

II. A Definition of Information Loss

When talking about the information loss induced by processing of signals, it is of prime importance to accompany this discussion by a well-based definition of information loss going beyond, but without lacking, intuition. Further, the definition shall also allow generalizations to stochastic processes and dynamical systems without contradicting previous statements. We try to meet both objectives with the following

Definition 1. Let X be an RV[1] on the samples space $\mathcal{X}$, and let Y be obtained by transforming X. We define the information loss induced by this transform as

$$L(X \to Y) = \sup_{\mathcal{P}} \Big(I(\hat{X};X) - I(\hat{X};Y) \Big) \tag{1}$$

where the supremum is over all partitions $\mathcal{P}$ of $\mathcal{X}$, and where $\hat{X}$ is obtained by quantizing X according to the partition $\mathcal{P}$ (see Fig. 1).

This Definition is motivated by the data processing inequality (cf. [3]), which states that the expression under the

[1]Note that X and all other involved RVs need not be scalar-valued.

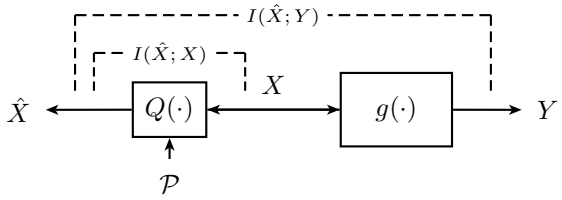

Fig. 1. Model for computing the information loss of a memoryless input-output system g. Q is a quantizer with partition $\mathcal{P}$.

supremum is always non-negative: Information loss is the worst-case reduction of information about $\hat{X}$ induced by transforming X. We now try to shed a little more light on Definition 1 in the following

Theorem 1. *The information loss of Definition 1 is given by:*

$$L(X \to Y) = \lim_{\hat{X} \to X} \Big(I(\hat{X};X) - I(\hat{X};Y) \Big) \tag{2}$$

$$= H(X|Y) \tag{3}$$

Proof: We start by noticing that

$$I(\hat{X};X) - I(\hat{X};Y) = H(\hat{X}|Y) \leq H(X|Y) \tag{4}$$

by the definition of mutual information and since both $\hat{X}$ and Y are functions of X. The inequality in (4) is due to data processing ($\hat{X}$ is a function of X). We now show that in the supremum over all partitions equality can be achieved.

To this end, observe that among all partitions of the sample space of X there is a sequence $\{\mathcal{P}_n\}$ of increasingly fine partitions[2] such that

$$\lim_{n \to \infty} \hat{X}_n = X \tag{5}$$

where $\hat{X}_n$ is the quantization of X induced by partition $\mathcal{P}_n$. By the axioms of entropy (e.g., [8, Ch. 14]), $H(\hat{X}_n|Y)$ is an increasing sequence in n with limit $H(X|Y)$. Thus, this limit represents the supremum in Definition 1, which proves (3).

Note further that each converging sequence $\hat{X} \to X$ contains a converging subsequence $\hat{X}_n \to X$ satisfying (5), and where $\hat{X}_{n+1}$ is obtained by refining the partition inducing $\hat{X}_n$. Therefore,

$$\lim_{\hat{X} \to X} H(\hat{X}|Y) = \lim_{n \to \infty} H(\hat{X}_n|Y) = H(X|Y) \tag{6}$$

which completes the proof. ∎

This Theorem shows that the supremum in Definition 1 is achieved for $\hat{X} \equiv X$, i.e., when we compute the difference between the *self-information* of the input and the information the output of the system contains about its input. This difference was shown to be identical to the conditional entropy of the input given the output – the quantity which is also used for quantifying the information loss due to noise or erasures (in the discrete-valued, memoryless case). In addition to that, the Theorem suggests a natural way to measure the information loss via measuring mutual informations, as it is depicted in Fig. 1. As we will see later (cf. Theorem 3), the considered partition does not have to be infinitely fine, but indeed a comparably coarse partition can deliver the correct result.

[2]i.e., $\mathcal{P}_{n+1}$ is a refinement of $\mathcal{P}_n$.

III. Problem Statement

Let $\mathbf{X} = [X_1, X_2, \ldots, X_N]$ be an N-dimensional RV with a probability measure $P_{\mathbf{X}}$ absolutely continuous w.r.t. the Lebesgue measure μ ($P_{\mathbf{X}} \ll \mu$). We require $P_{\mathbf{X}}$ to be concentrated on $\mathcal{X} \subseteq \mathbb{R}^N$. This RV, which possesses a unique probability density function (PDF) $f_{\mathbf{X}}$, is the input to the following multivariate, vector-valued function:

Definition 2. Let $\mathbf{g}$: $\mathcal{X} \to \mathcal{Y}$, $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^N$, be a surjective, Borel-measurable function defined in a piecewise manner:

$$\mathbf{g}(\mathbf{x}) = \begin{cases} \mathbf{g}_1(\mathbf{x}), & \text{if } \mathbf{x} \in \mathcal{X}_1 \\ \mathbf{g}_2(\mathbf{x}), & \text{if } \mathbf{x} \in \mathcal{X}_2 \\ \vdots \end{cases} \tag{7}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_N]$ and $\mathbf{g}_i$: $\mathcal{X}_i \to \mathcal{Y}_i$ bijectively[3]. Furthermore, let the Jacobian matrix $\mathcal{J}_{\mathbf{g}}(\cdot)$ exist on the closures of $\mathcal{X}_i$. In addition to that, we require the Jacobian determinant, $|\det \mathcal{J}_{\mathbf{g}}(\cdot)|$, to be non-zero $P_{\mathbf{X}}$-almost everywhere.

In accordance with previous work [7] the $\mathcal{X}_i$ are disjoint sets of positive $P_{\mathbf{X}}$-measure which unite to $\mathcal{X}$, i.e., $\bigcup_i \mathcal{X}_i = \mathcal{X}$ and $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$ if $i \neq j$. Clearly, also the $\mathcal{Y}_i$ unite to $\mathcal{Y}$, but need not be disjoint. This definition ensures that the preimage $\mathbf{g}^{-1}[\mathbf{y}]$ of each element $\mathbf{y} \in \mathcal{Y}$ is a countable set.

Using the method of transformation [8, pp. 244] one obtains the PDF of the N-dimensional output RV $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]$ as

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{\mathbf{x}_i \in \mathbf{g}^{-1}[\mathbf{y}]} \frac{f_{\mathbf{X}}(\mathbf{x}_i)}{|\det \mathcal{J}_{\mathbf{g}}(\mathbf{x}_i)|} \tag{8}$$

where the sum is over all elements of the preimage. Note that since $\mathbf{Y}$ possesses a density, the corresponding probability measure $P_{\mathbf{Y}}$ is also absolutely continuous w.r.t. the Lebesgue measure.

IV. Main Results

We now state our main results:

Theorem 2. *The information loss induced by a function* $\mathbf{g}$ *satisfying Definition 2 is given as*

$$H(\mathbf{X}|\mathbf{Y}) = \int_{\mathcal{X}} f_{\mathbf{X}}(\mathbf{x}) \log \left(\frac{\sum_{\mathbf{x}_i \in \mathbf{g}^{-1}[\mathbf{g}(\mathbf{x})]} \frac{f_{\mathbf{X}}(\mathbf{x}_i)}{|\det \mathcal{J}_{\mathbf{g}}(\mathbf{x}_i)|}}{\frac{f_{\mathbf{X}}(\mathbf{x})}{|\det \mathcal{J}_{\mathbf{g}}(\mathbf{x})|}} \right) d\mathbf{x}. \tag{9}$$

The proof of this Theorem can be found in the Appendix and, in a modified version for univariate functions, in [7]. Note that for univariate functions the Jacobian determinant is replaced by the derivative of the function.

[3]In the univariate case, i.e., for $N = 1$, this is equivalent to requiring that g is piecewise strictly monotone.

Corollary 1. *The information loss induced by a function* $\mathbf{g}$ *satisfying Definition 2 is given as*

$$H(\mathbf{X}|\mathbf{Y}) = h(\mathbf{X}) - h(\mathbf{Y}) + \mathrm{E}\left\{\log|\det\mathcal{J}_{\mathbf{g}}(\mathbf{X})|\right\} \tag{10}$$

Proof: The proof is obtained by recognizing the PDF of $\mathbf{Y}$ inside the logarithm in (9) and by splitting the logarithm. ■

This result is particularily interesting because it provides a link between information loss and differential entropies already anticipated in [8, pp. 660]. There, it was claimed that

$$h(\mathbf{Y}) \leq h(\mathbf{X}) + \mathrm{E}\left\{\log|\det\mathcal{J}_{\mathbf{g}}(\mathbf{X})|\right\} \tag{11}$$

where equality holds iff $\mathbf{g}$ is bijective. While (11) is actually another version of the DPI, Corollary 1 quantifies how much information is lost by processing. In addition to that, a very similar expression denoted as *folding entropy* has been presented in [9], although in a completely different setting analyzing the entropy production of *autonomous* dynamical systems.

We now introduce a discrete RV W which depends on the set $\mathcal{X}_i$ from which $\mathbf{X}$ was taken. In other words, for all i we have $W = w_i$ iff $\mathbf{x} \in \mathcal{X}_i$. One can interpret this RV as being generated by a vector quantization of $\mathbf{X}$ with a partition $\mathcal{P} = \{\mathcal{X}_i\}$. With this new RV we can state

Theorem 3. *The information loss is identical to the uncertainty about the set* $\mathcal{X}_i$ *from which the input was taken, i.e.,*

$$H(\mathbf{X}|\mathbf{Y}) = H(W|\mathbf{Y}). \tag{12}$$

The proof follows closely the proof provided in [7] and thus is omitted. However, this equivalence suggests a way of measuring information loss by means of proper quantization: Since $H(W|\mathbf{Y}) = I(W;\mathbf{X}) - I(W;\mathbf{Y})$ the loss can be determined by measuring mutual informations, which in this case are always finite (or, at least, bounded by $H(W)$). In contrary to that, the mutual information in (2) of Theorem 1 diverge to infinity; This expression was used in [7] for the information loss, highlighting the fact that both the self-information of $\mathbf{X}$ and the information transfer from $\mathbf{X}$ to $\mathbf{Y}$ are infinite.

The interpretation derived from Theorem 3 allows us now to provide upper bounds on the information loss:

Theorem 4. *The information loss is upper bounded by*

$$\begin{aligned} H(\mathbf{X}|\mathbf{Y}) &\leq \int_{\mathcal{Y}} f_{\mathbf{Y}}(\mathbf{y}) \log|\mathbf{g}^{-1}[\mathbf{y}]| d\mathbf{y} && (13)\\ &\leq \log\left(\sum_i \int_{\mathcal{Y}_i} f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}\right) && (14)\\ &\leq \max_{\mathbf{y}} \log|\mathbf{g}^{-1}[\mathbf{y}]|. && (15) \end{aligned}$$

Proof: We give here only a sketch of the proof: The first inequality results from bounding $H(W|\mathbf{Y}=\mathbf{y})$ by the entropy of a uniform distribution on the preimage of $\mathbf{y}$. Jensen's inequality yields the second line of the Theorem. The coarsest bound is obtained by replacing the cardinality of the preimage by its maximal value. ■

In this Theorem, we bounded the information loss given a certain output by the cardinality of the preimage. While the first bound considers the fact that the cardinality may actually depend on the output itself, the last bound incorporates the maximum cardinality only. In cases where the function from Definition 2 is defined not on a countable but on a finite number of subdomains this finite number can act as an upper bound (cf. [7]). Another straightforward upper bound, which is independent from the bounds in Theorem 4 is obtained from Theorem 3 by removing conditioning:

$$H(\mathbf{X}|\mathbf{Y}) \leq H(W) = -\sum_i p_i \log p_i \tag{16}$$

where $p_i = P_{\mathbf{X}}(\mathcal{X}_i) = \int_{\mathcal{X}_i} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$. It has to be noted, though, that depending on the function $\mathbf{g}$ all these bounds can be infinite while the information loss remains finite.

A further implication of introducing this discrete RV W is that it allows us to perform investigations about reconstructing the input from the output. Currently, a Fano-type inequality bounding the reconstruction error by the information loss is under development. In addition to that, new upper bounds on the information loss related to the reconstruction error of optimal (in the *maximum a posteriori* sense) and of simpler, sub-optimal estimators are analyzed.

V. Functions with Infinite Information Loss

We now drop the requirement of local bijectivity in Definition 2 to analyze a wider class of surjective, Borel-measurable functions $\mathbf{g}$: $\mathcal{X} \to \mathcal{Y}$. We keep the requirement that $P_{\mathbf{X}} \ll \mu$ and thus $\mathbf{X}$ possesses a density $f_{\mathbf{X}}$ (positive on $\mathcal{X}$ and zero elsewhere). We maintain

Theorem 5. *Let* $\mathbf{g}$: $\mathcal{X} \to \mathcal{Y}$ *be a Borel-measurable function and let the continuous RV* $\mathbf{X}$ *be the input to this function. If there exists a set* $B \subseteq \mathcal{Y}$ *of positive* $P_{\mathbf{Y}}$*-measure such that the preimage* $\mathbf{g}^{-1}[\mathbf{y}]$ *is uncountable for every* $\mathbf{y} \in B$*, then the information loss is infinite.*

Proof: We notice that since $B \subseteq \mathcal{Y}$

$$\begin{aligned} H(\mathbf{X}|\mathbf{Y}) &= \int_{\mathcal{Y}} H(\mathbf{X}|\mathbf{Y}=\mathbf{y}) dP_{\mathbf{Y}}(\mathbf{y}) && (17)\\ &\geq \int_{B} H(\mathbf{X}|\mathbf{Y}=\mathbf{y}) dP_{\mathbf{Y}}(\mathbf{y}) && (18) \end{aligned}$$

where the integrals are now written as Lebesgue integrals, since $P_{\mathbf{Y}}$ now not necessarily possesses a density.

Since on B the preimage of every element is uncountable, we obtain with [4] and the references therein $H(\mathbf{X}|\mathbf{Y}=\mathbf{y}) = \infty$ for all $\mathbf{y} \in B$, and, thus, $H(\mathbf{X}|\mathbf{Y}) = \infty$. ■

Note that the requirement of B being a set of positive $P_{\mathbf{Y}}$-measure cannot be dropped, as Example 4 in Section VI illustrates. We immediately obtain the following

Corollary 2. *Let* $\mathbf{g}$: $\mathcal{X} \to \mathcal{Y}$ *be a Borel-measurable function and let the continuous RV* $\mathbf{X}$ *be the input to this function. If the probability measure of the output,* $P_{\mathbf{Y}}$*, possesses a non-vanishing discrete component, the information loss is infinite.*

Proof: According to the Lebesgue-Radon-Nikodym theorem [10, pp. 121] every measure can be decomposed in a component absolutely continuous w.r.t. μ and a component singular to μ. The latter part can further be decomposed into a singular continuous and a discrete part, where the latter places positive $P_{\mathbf{Y}}$-mass on points. Let $\mathbf{y}^*$ be such a point, i.e., $P_{\mathbf{Y}}(\mathbf{y}^*) > 0$. As an immediate consequence, $P_{\mathbf{X}}(\mathbf{g}^{-1}[\mathbf{y}^*]) > 0$, which is only possible if $\mathbf{g}^{-1}[\mathbf{y}^*]$ is uncountable ($P_{\mathbf{X}} \ll \mu$). ■

This result is also in accordance with intuition, as the analysis of a simple quantizer shows: While the entropy of the input RV is infinite ($I(\hat{\mathbf{X}};\mathbf{X}) \to \infty$ for $\hat{\mathbf{X}} \to \mathbf{X}$; cf. [8, pp. 654]), the quantized output can contain only a finite amount of information ($I(\hat{\mathbf{X}};\mathbf{Y}) \to H(\mathbf{Y}) < \infty$). In addition to that, the preimage of each possible output value $\mathbf{y}$ is a set of positive $P_{\mathbf{X}}$-measure. The loss, as a consequence, is infinite.

While for the quantizer the preimage of each possible output value is a set of positive measure, there certainly are functions for which some outputs have a countable preimage and some whose preimage is a non-null set. An example of such a system is the limiter [8, Ex. 5-4]. For such systems it can be shown that both the information loss $L(\mathbf{X} \to \mathbf{Y}) = H(\mathbf{X}|\mathbf{Y})$ and the information transfer $I(\mathbf{X};\mathbf{Y})$ are infinite.

Finally, there exist functions $\mathbf{g}$ for which the preimages of all output values $\mathbf{y}$ are null sets, but which still fulfill the conditions of Theorem 5. Functions which project $\mathcal{X}$ on a lower-dimensional subspace of $\mathbb{R}^N$ fall into that category.

VI. Examples

In this Section we illustrate our theoretical results with the help of examples. The logarithm is taken to base 2 unless otherwise noted.

A. Example 1: A two-dimensional transform with finite information loss

Let $\mathbf{X}$ be uniformly distributed on the square $\mathcal{X} = [-a,a] \times [-a,a]$. Equivalently, the two constituing RVs X_1 and X_2 are independent and uniformly distributed on $[-a,a]$. In other words, while $f_{\mathbf{X}}(\mathbf{x}) = 1/4a^2$ for all $\mathbf{x} \in \mathcal{X}$, we have $f_X(x_i) = 1/2a$ for $x_i \in [-a,a]$ and $i = 1,2$.

We consider a function $\mathbf{g}$ performing the mapping:

$$Y_1 = X_1 \tag{19}$$
$$Y_2 = |X_1 - X_2| \tag{20}$$

The corresponding Jacobian matrix is a triangular matrix

$$\mathcal{J}_{\mathbf{g}}(\mathbf{x}) = \begin{bmatrix} 1 & 0 \\ \operatorname{sgn}(x_1 - x_2) & \operatorname{sgn}(x_2 - x_1) \end{bmatrix} \tag{21}$$

where $\operatorname{sgn}(\cdot)$ is the sign-function. From this immediately follows that the magnitude of the determinant of the Jacobian matrix is unity for all possible values of $\mathbf{X}$, i.e., $|\det \mathcal{J}_{\mathbf{g}}(\mathbf{x})| = 1$ for all $\mathbf{x} \in \mathcal{X}$. The subsets of $\mathcal{X}$ on which the partitioned functions $\mathbf{g}_i$ are bijective are no intervals in this case; they are the triangular halves of the square induced by $x_1 = x_2$ (see Fig. 2)

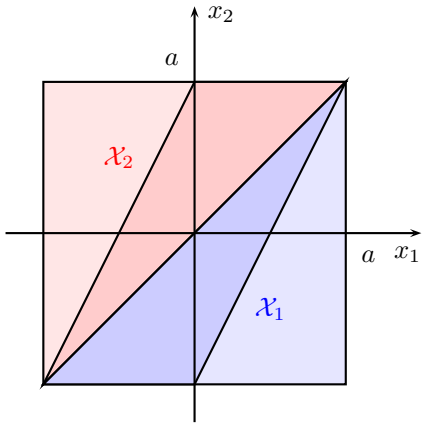

Fig. 2. Subdomains of Example 1. The partitioned functions $\mathbf{g}_i$ restricted to the domain of either color are bijective. Furthermore, the overall function $\mathbf{g}$ is bijective in areas with light shading.

$$\mathcal{X}_1 = \{[x_1,x_2] \in \mathcal{X} : \ x_1 > x_2\} \tag{22}$$
$$\mathcal{X}_2 = \{[x_1,x_2] \in \mathcal{X} : \ x_1 \leq x_2\}. \tag{23}$$

The preimage of $\mathbf{g}(\mathbf{x})$ is, in any case,

$$\{[x_1,x_2],[x_1,2x_1 - x_2]\} \cap \mathcal{X}. \tag{24}$$

The transform $\mathbf{g}$ is bijective whenever $[x_1, 2x_1 - x_2] \notin \mathcal{X}$, i.e., if $|2x_1 - x_2| > a$.

With the PDF of $\mathbf{X}$ and of its components we obtain for the information loss

$$H(\mathbf{X}|\mathbf{Y}) = \int_{-a}^{a}\int_{-a}^{a} \frac{1}{4a^2} \log\left(\frac{\frac{1}{2a} + f_X(2x_1 - x_2)}{\frac{1}{2a}}\right) dx_1 dx_2 \tag{25}$$

which is non-zero only if $-a \leq 2x_1 - x_2 \leq a$ (numerator and denominator cancel otherwise; no loss occurs in the bijective domain of the function). As a consequence,

$$H(\mathbf{X}|\mathbf{Y}) = \int_{-a}^{a}\int_{\frac{x_2-a}{2}}^{\frac{x_2+a}{2}} \frac{\log 2}{4a^2} dx_1 dx_2 \tag{26}$$
$$= \int_{-a}^{a} \frac{1}{4a} dx_2 = \frac{1}{2}. \tag{27}$$

The information loss is identical to a half bit. This is intuitive when looking at Fig. 2, where it can be seen that any information loss occurs only on one half of the domain $\mathcal{X}$ (shaded in stronger colors). By destroying the sign information, in this area the information loss is equal to one bit.

B. Example 2: Squaring a Gaussian RV

Let X be a zero-mean Gaussian RV with unit variance and differential entropy $h(X) = \frac{1}{2}\ln(2\pi e)$ measured in nats. We consider the square of this RV, $Y = g(X) = X^2$, to illustrate the connection between information loss and

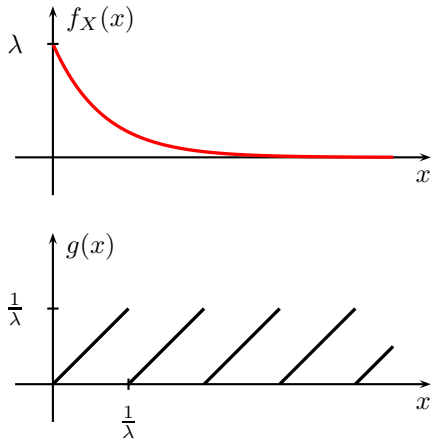

Fig. 3. PDF f_X and piecewise linear function g of Example 3.

differential entropy. The square of a Gaussian RV, Y, is χ^2-distributed with one degree of freedom. Thus, the differential entropy of Y is given by [11]

$$h(Y) = \frac{1}{2} + \ln\left(2\Gamma\left(\frac{1}{2}\right)\right) + \frac{1}{2}\psi\left(\frac{1}{2}\right) \tag{28}$$

$$= \frac{1}{2} + \frac{1}{2}\ln\pi - \frac{\gamma}{2} \tag{29}$$

where $\Gamma(\cdot)$ and $\psi(\cdot)$ are the gamma- and digamma-functions [12, Ch. 6] and γ is the Euler-Mascheroni constant [12, pp. 3]. With some calculus we obtain for the expected value of the derivative (taking the place of the Jacobian determinant in the univariate case)

$$\mathrm{E}\left\{\ln|2x|\right\} = \frac{1}{2}\ln 2 - \frac{\gamma}{2}. \tag{30}$$

Subtracting differential entropies and adding the expected value of the derivative yields the information loss

$$H(X|Y) = h(X) - h(Y) + \mathrm{E}\left\{\ln|2x|\right\} \tag{31}$$

$$= \frac{1}{2}\ln(2\pi e) - \frac{1}{2} - \frac{1}{2}\ln(\pi) + \frac{1}{2}\ln 2 \tag{32}$$

$$= \ln 2 \tag{33}$$

again measured in nats. Changing the base of the logarithm to 2 we obtain an information loss of one bit. This is in perfect accordance with a previous result showing that the information loss of a square-law device is equal to one bit if the PDF of the input has even symmetry [7].

C. Example 3: Exponential RV and infinite bounds

In this example we consider an exponential input with PDF

$$f_X(x) = \lambda \mathrm{e}^{-\lambda x} \tag{34}$$

and a piecewise linear function

$$g(x) = x - \frac{\lfloor \lambda x \rfloor}{\lambda}. \tag{35}$$

The PDF and the function are depicted in Fig. 3.

We obviously have $\mathcal{X} = [0, \infty)$ and $\mathcal{Y} = [0, \frac{1}{\lambda})$, while g partitions $\mathcal{X}$ in a countable number of intervals of length $\frac{1}{\lambda}$. In other words,

$$\mathcal{X}_k = \left[\frac{k-1}{\lambda}, \frac{k}{\lambda}\right) \tag{36}$$

and $g(\mathcal{X}_k) = \mathcal{Y}$ for all $k = 1, 2, \ldots$. From this follows that for every $y \in \mathcal{Y}$ the preimage contains an element from each subdomain $\mathcal{X}_k$; thus, the bounds from Theorem 4 all evaluate to $H(X|Y) \leq \infty$. However, it can be shown that the other bound, $H(X|Y) \leq H(W)$ is tight in this case: With

$$p_k = P_X(\mathcal{X}_k) = \int_{\mathcal{X}_k} f_X(x)dx = (1 - \mathrm{e}^{-1})\mathrm{e}^{-k+1} \tag{37}$$

we obtain $H(W) = -\log(1-\mathrm{e}^{-1}) + \frac{\mathrm{e}^{-1}}{1-\mathrm{e}^{-1}} \approx 1.24$. The same result is obtained for a direct evaluation of Theorem 2.

D. Example 4: An almost invertible transform with zero information loss

As a next example consider a two-dimensional RV $\mathbf{X}$ which places probability mass uniformly on the unit disc, i.e.,

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \frac{1}{\pi}, & \text{if } ||\mathbf{x}|| \leq 1 \\ 0, & \text{else} \end{cases} \tag{38}$$

where $||\cdot||$ is the Euclidean norm. Thus, $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^2 : ||\mathbf{x}|| \leq 1\}$. The cartesian coordinates $\mathbf{x}$ are now transformed to polar coordinates in a special way, namely:

$$y_1 = \begin{cases} ||\mathbf{x}||, & \text{if } ||\mathbf{x}|| < 1 \\ 0, & \text{else} \end{cases} \tag{39}$$

$$y_2 = \begin{cases} \arctan(\frac{x_2}{x_1}) + \pi(1 - \operatorname{sgn}(x_1)), & \text{if } 0 < ||\mathbf{x}|| < 1 \\ 0, & \text{else} \end{cases} \tag{40}$$

This mapping together with the domains of $\mathbf{X}$ and $\mathbf{Y}$ is illustrated in Fig. 4 (left and upper right diagram).

As a direct consequence we have $\mathcal{Y} = (0,1) \times [0, 2\pi) \cup \{0,0\}$. Observe that not only the point $\mathbf{x} = \{0,0\}$ is mapped to the point $\mathbf{y} = \{0,0\}$, but that also the unit circle $\mathcal{S} = \{\mathbf{x} : ||\mathbf{x}|| = 1\}$ is mapped to $\mathbf{y} = \{0,0\}$. As a consequence, the preimage of $\{0,0\}$ under $\mathbf{g}$ is uncountable. However, since a circle in $\mathbb{R}^2$ is a Lebesgue null-set and thus $P_{\mathbf{X}}(\mathcal{S}) = 0$, also $P_{\mathbf{Y}}(\{0,0\}) = 0$ and the conditions of Theorem 5 are not met. Indeed, since $H(\mathbf{X}|\mathbf{Y} = \mathbf{y}) = 0$ $P_{\mathbf{Y}}$-almost everywhere, it can be shown that $H(\mathbf{X}|\mathbf{Y}) = 0$.

E. Example 5: A mapping to a subspace of lower dimensionality

Consider again a uniform distribution on the unit disc, as it was used in Example 4. Now, however, let $\mathbf{g}$ be such that only the radius is computed while the angle is lost, i.e.,

$$y_1 = ||\mathbf{x}|| \tag{41}$$

$$y_2 = 0. \tag{42}$$

Note that here only the origin $\{0,0\}$ is mapped bijectively, while for all other $\mathbf{y} \in \mathcal{Y} = [0,1] \times \{0\}$ the preimage under

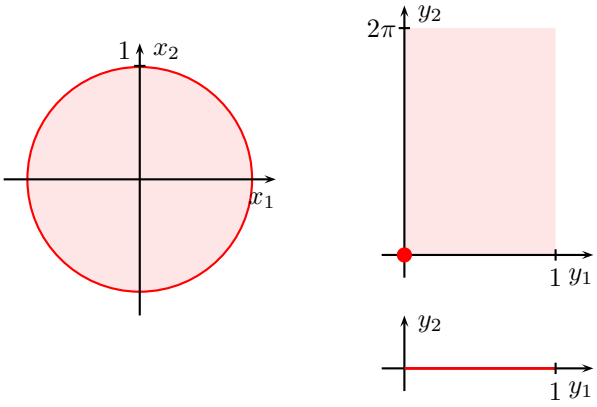

Fig. 4. Mapping of domains in Examples 4 and 5. The solid red circle in the left diagram and the red dot in the upper right diagram correspond to each other, illustrating the mapping of an uncountable $P_{\mathbf{X}}$-null set to a point. The lightly shaded areas are mapped bijectively in Example 4. In Example 5, the disc in the left diagram is mapped to the solid red line in the lower right diagram.

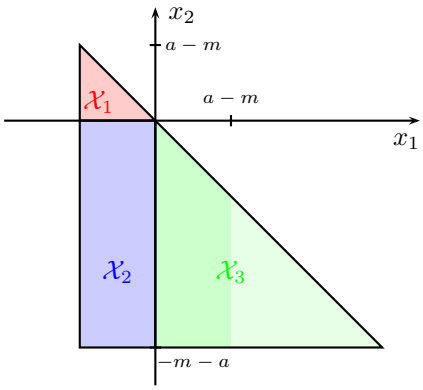

Fig. 5. Subdomains of Example 6. The functions $\mathbf{g}_i$ restricted to a domain of either color are bijective. Furthermore, the overall function $\mathbf{g}$ is bijective in areas with light shading.

$\mathbf{g}$ is uncountable (a circle around the origin with radius y_1). Indeed, in this particular example, the probability measure $P_{\mathbf{Y}}$ is *not* discrete, but singular continuous: Each point has zero $P_{\mathbf{Y}}$-measure (circles are Lebesgue null-sets), but $P_{\mathbf{Y}}$ is not absolutely continuous w.r.t. the two-dimensional Lebesgue measure μ. Clearly, $\mu(\mathcal{Y}) = 0$ while $P_{\mathbf{Y}}(\mathcal{Y}) = 1$. Since the preimage is uncountable on a set of positive $P_{\mathbf{Y}}$-measure, we have $H(\mathbf{X}|\mathbf{Y}) = \infty$.

F. Example 6: Another two-dimensional transform with finite information loss

Finally, consider a uniform distribution on a triangle defined by

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^2 : x_1 \in [m-a, m+a], x_2 \in [-m-a, -x_1]\} \tag{43}$$

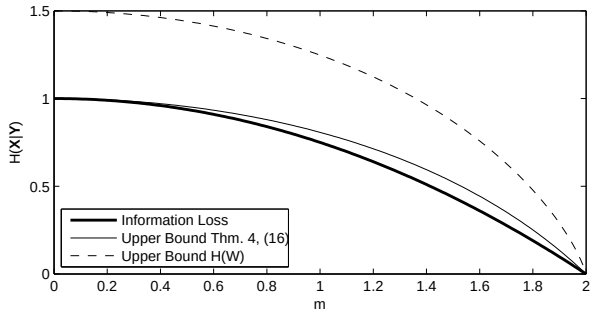

Fig. 6. Information loss $H(\mathbf{X}|\mathbf{Y})$ of Example 6 for $a = 2$.

where $0 \leq m \leq a$ and $a > 0$. Thus, the PDF of $\mathbf{X}$ is given as $f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2a^2}$ if $\mathbf{x} \in \mathcal{X}$ and zero elsewhere (see Fig. 5). The function $\mathbf{g}$ takes the magnitude of each coordinate, i.e., $y_i = |x_i|$, where $i = 1, 2$. We now try to derive the information loss as a function of m.

First we can identify three subsets of $\mathcal{X}$ which are mapped bijectively by restricting $\mathbf{g}$ to these sets, namely $\mathcal{X}_1 = \{\mathbf{x} \in \mathcal{X} : x_1 \leq 0, x_2 \geq 0\}$, $\mathcal{X}_2 = \{\mathbf{x} \in \mathcal{X} : x_1 \leq 0, x_2 < 0\}$, and $\mathcal{X}_3 = \{\mathbf{x} \in \mathcal{X} : x_1 > 0, x_2 < 0\}$. Furthermore, for $m \geq 0$ a part of $\mathcal{X}_3$ is mapped bijectively by $\mathbf{g}$ (lighter shading in Fig. 5). The probability mass contained in this subset $\mathcal{X}_b$ can be shown to equal $P_b = P_{\mathbf{X}}(\mathcal{X}_b) = \frac{m^2}{a^2}$. For all other possible input values $\mathbf{x}$ the preimage of $\mathbf{g}(\mathbf{x})$ has exactly two elements: One of them is located in $\mathcal{X}_2$, the other either in $\mathcal{X}_1$ or in $\mathcal{X}_3 \setminus \mathcal{X}_b$. Due to the uniformity of $\mathbf{X}$ and since the Jacobian determinant is identical to unity for all $\mathbf{x} \in \mathcal{X}$ both of these preimages are equally likely. Thus, on $\mathcal{X} \setminus \mathcal{X}_b$ the information loss is identical to one bit. In other words,

$$H(\mathbf{X}|\mathbf{Y} = \mathbf{y}) = 1 \tag{44}$$

for all $\mathbf{y} \in \mathbf{g}(\mathcal{X} \setminus \mathcal{X}_b)$. We therefore obtain with $P_b = \frac{m^2}{a^2}$ an information loss equal to $H(\mathbf{X}|\mathbf{Y}) = 1 - \frac{m^2}{a^2}$.

From the probability masses contained in the sets $\mathcal{X}_1$, $\mathcal{X}_2$, and $\mathcal{X}_3$ we can compute an upper bound on the information loss:

$$H(W) = \frac{m^2}{2a^2} + \frac{3}{2} - \log\frac{a^2 - m^2}{a^2} + \frac{m}{a}\log\frac{a-m}{a+m}. \tag{45}$$

And evaluating the bounds of Theorem 4 yields

$$H(\mathbf{X}|\mathbf{Y}) \leq 1 - \frac{m^2}{a^2} \leq \log(2 - \frac{m^2}{a^2}) \leq 1 \tag{46}$$

which for $m = 0$ all reduce to one bit. In particular, it can be seen that in this case the smallest bound of Theorem 4 is exact.

The exact information loss, together with the second smallest bound from Theorem 4 and with the bound from $H(W)$, is shown in Fig. 6. As it can be seen, the closer the parameter m approaches a, the smaller the information loss gets. Conversely, for $m = 0$ the information loss is exactly one bit. Moreover, it turns out that the bound from $H(W)$ is rather loose in this case.

VII. Conclusion

In this work, we proposed a mathematically concise definition of information loss for the purpose of establishing a system theory from an information-theoretic point of view. For a certain class of multivariate, vector-valued functions and continuous input variables this information loss was quantified, and the result is accompanied by convenient upper bounds. We further showed a connection between information loss and the differential entropies of the input and output variables.

Finally, a class of systems has been identified for which the information loss is infinite. Vector-quantizers and limiters belong to that class, but also functions which project the input space onto a space of lower dimensionality.

Appendix

Proof of Theorem 2

For the proof we use (2) of Theorem 2, where we take the limit of a sequence of increasingly fine partitions $\mathcal{P}_n = \{\hat{\mathcal{X}}_k^{(n)}\}$ satisfying (5). For a given n we write the resulting mutual information $I(\hat{\mathbf{X}}_n;\mathbf{X})$ as

$$I(\hat{\mathbf{X}}_n;\mathbf{X}) = \mathrm{E}\left\{D(f_{\mathbf{X}|\hat{\mathbf{X}}_n}(\cdot,\hat{\mathbf{x}})||f_{\mathbf{X}}(\cdot))\right\} \tag{47}$$

where $D(\cdot||\cdot)$ denotes the Kullback-Leibler divergence and the expection is w.r.t. $\hat{\mathbf{X}}_n$. Note that for each possible outcome $\hat{\mathbf{x}}_k$ of $\hat{\mathbf{X}}_n$ the conditional probability measure $P_{\mathbf{X}|\hat{\mathbf{x}}_k}$ is absolutely continuous w.r.t. the Lebesgue measure (cf. Section V). It thus possesses a density

$$f_{\mathbf{X}|\hat{\mathbf{X}}_n}(\mathbf{x},\hat{\mathbf{x}}_k) = \begin{cases} \frac{f_{\mathbf{X}}(\mathbf{x})}{p(\hat{\mathbf{x}}_k)}, & \text{if } \mathbf{x}\in\hat{\mathcal{X}}_k^{(n)} \\ 0, & \text{else} \end{cases} \tag{48}$$

where $p(\hat{\mathbf{x}}_k) = P_{\mathbf{X}}(\hat{\mathcal{X}}_k^{(n)})$. With the definition of the Kullback-Leibler divergence [3, Lemma 5.2.3] and [8, Thm. 5-1] we can write the difference of mutual informations in Theorem 1 as

$$I(\hat{\mathbf{X}}_n;\mathbf{X}) - I(\hat{\mathbf{X}}_n;\mathbf{Y}) = \sum_k p(\hat{\mathbf{x}}_k)\int_{\hat{\mathcal{X}}_k^{(n)}} \frac{f_{\mathbf{X}}(\mathbf{x})}{p(\hat{\mathbf{x}}_k)} \log\left(\frac{f_{\mathbf{X}|\hat{\mathbf{X}}_n}(\mathbf{x},\hat{\mathbf{x}}_k) f_{\mathbf{Y}}(\mathbf{g}(\mathbf{x}))}{f_{\mathbf{Y}|\hat{\mathbf{X}}_n}(\mathbf{g}(\mathbf{x}),\hat{\mathbf{x}}_k) f_{\mathbf{X}}(\mathbf{x})}\right) d\mathbf{x}. \tag{49}$$

Rewriting with the indicator function

$$\mathbf{I}_A(x) = \begin{cases} 1, & \text{if } x\in A \\ 0, & \text{else} \end{cases} \tag{50}$$

this yields

$$I(\hat{\mathbf{X}}_n;\mathbf{X}) - I(\hat{\mathbf{X}}_n;\mathbf{Y}) = \int_{\mathcal{X}} f_{\mathbf{X}}(\mathbf{x}) \sum_k \left(\mathbf{I}_{\hat{\mathcal{X}}_k^{(n)}}(\mathbf{x}) \log\left(\frac{f_{\mathbf{X}|\hat{\mathbf{X}}_n}(\mathbf{x},\hat{\mathbf{x}}_k) f_{\mathbf{Y}}(\mathbf{g}(\mathbf{x}))}{f_{\mathbf{Y}|\hat{\mathbf{X}}_n}(\mathbf{g}(\mathbf{x}),\hat{\mathbf{x}}_k) f_{\mathbf{X}}(\mathbf{x})}\right)\right) d\mathbf{x}.$$

We can now exploit the relationship (8) for the conditional PDF of $\mathbf{Y}$ given $\hat{\mathbf{X}}_n$, and with (48) we realize that the function under the integral is monotonically increasing in n: Indeed, for finer partitions it is less likely that any element of the preimage $\mathbf{g}^{-1}[\mathbf{g}(\mathbf{x})]$ other than $\mathbf{x}$ lies in $\hat{\mathcal{X}}_k^{(n)}$, thus $f_{\mathbf{Y}|\hat{\mathbf{X}}_n}(\mathbf{g}(\mathbf{x}),\hat{\mathbf{x}}_k)$ converges to $\frac{f_{\mathbf{X}|\hat{\mathbf{X}}_n}(\mathbf{x},\hat{\mathbf{x}}_k)}{|\det\mathcal{J}_{\mathbf{g}}(\mathbf{x})|}$. This holds for all k, thus, invoking the monotone converge theorem [10, pp. 21] and cancelling the conditional PDFs eliminates the dependence on k and the sum over indicator functions ($\bigcup_k \hat{\mathcal{X}}_k^{(n)} = \mathcal{X}$). Substituting the PDF of $\mathbf{Y}$ with (8) completes the proof. ■

References

[1] T. S. Han, "Musing upon information theory," XXXI Shannon Lecture, 2010, presented at IEEE Int. Sym. on Information Theory (ISIT).

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, pp. 379–423, 623–656, Oct. 1948.

[3] R. M. Gray, *Entropy and Information Theory*. New York, NY: Springer, 1990.

[4] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco: Holden Day, 1964.

[5] S. Watanabe and C. T. Abraham, "Loss and recovery of information by coarse observation of stochastic chain," *Information and Control*, vol. 3, no. 3, pp. 248–278, Sep. 1960.

[6] B. C. Geiger and G. Kubin, "Some results on the information loss in dynamical systems," in *Proc. IEEE Int. Sym. Wireless Communication Systems (ISWSC)*, Aachen, Nov. 2011, accepted; preprint available: `arXiv:1106.2404 [cs.IT]`.

[7] B. C. Geiger, C. Feldbauer, and G. Kubin, "Information loss in static nonlinearities," in *Proc. IEEE Int. Sym. Wireless Communication Systems (ISWSC)*, Aachen, Nov. 2011, accepted; preprint available: `arXiv:1102.4794 [cs.IT]`.

[8] A. Papoulis and U. S. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. New York, NY: McGraw Hill, 2002.

[9] D. Ruelle, "Positiviy of entropy production in nonequilibrium statistical mechanics," *J.Stat.Phys.*, vol. 85, pp. 1–23, 1996.

[10] W. Rudin, *Real and Complex Analysis*, 3rd ed. New York, NY: McGraw-Hill, 1987.

[11] A. C. Verdugo Lazo and P. N. Rathie, "On the entropy of continuous probability distributions," *IEEE Transactions on Information Theory*, vol. IT-24, pp. 120–122, 1978.

[12] M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th ed. Dover Publications, 1972.